# Evaluating permutation-based inference for partial least squares analysis of neuroimaging data

Matthew Danyluik[1,2*], Yashar Zeighami[1,3], Alice Mukora[1,2], Martin Lepage[3,4,5], Jai Shah[3,4], Ridha Joober[3,4], Bratislav Misic[6,7], Yasser Iturria-Medina[6,7,8,9*†], M. Mallar Chakravarty[1,3,9*†]

[1] Cerebral Imaging Centre, Douglas Mental Health University Institute, Montréal, Canada
[2] Integrated Program in Neuroscience, McGill University, Montréal, Canada
[3] Department of Psychiatry, McGill University, Montréal, Canada
[4] Prevention and Early Intervention for Psychosis, Douglas Mental Health University Institute, Montréal, Canada
[5] Department of Psychology, McGill University, Montréal, Canada
[6] McConnell Brain Imaging Centre, McGill University, Montréal, Canada
[7] Department of Neurology and Neurosurgery, McGill University, Montréal, Canada
[8] Ludmer Centre for Neuroinformatics and Mental Health, McGill University, Montréal, Canada
[9] Department of Biological and Biomedical Engineering, McGill University, Montréal, Canada
* Corresponding authors.
Contact: matthew.danyluik@mail.mcgill.ca (MD), yasser.itturiamedina@mcgill.ca (YIM), mallar.chakravarty@mcgill.ca (MMC)
† These authors contributed equally to this work.

## Abstract

Partial least squares (PLS) is actively leveraged in neuroimaging work, typically to map latent variables (LVs) representing brain-behaviour associations. Canonically, LVs are considered statistically significant if they tend to capture more covariance than LVs derived from permuted data, with a Procrustes rotation applied to map each set of permuted LVs to the space defined by the originals, creating an "apples to apples" comparison. Yet, it has not been established whether applying the rotation makes the permutation test more sensitive to the true LVs in a dataset, and it is unclear if significant LVs can be drawn from data with no meaningful between-feature covariance. Accordingly, we performed PLS analyses across a range of simulated datasets with known latent effects, observing that the Procrustes rotation systematically weakened the null distributions for the first LV. By extension, the first LV was nearly always significant, regardless of whether the effect was weak, undersampled, noisy, or simulated at all. But, if no rotation was applied, all possible LVs tended to be significant as we increased the sample size of UK BioBank datasets. Meanwhile, LV strength and stability metrics accurately tracked our confidence that effects were present in simulated data, and allowed for a more nuanced assessment of which LVs may be relevant in the UK BioBank. We end by presenting a list of considerations for researchers implementing PLS permutation testing, and by discussing promising alternative tests which may alleviate the concerns raised by our findings.

## Keywords

Partial least squares, permutation testing, statistical inference, multivariate analysis

## 1. Background

The partial least squares (PLS) correlation is a popular technique in neuroimaging research, typically used to identify sets of brain and behavioural measures that covary across a sample, referred to as latent variables (LVs) (Abdi et al., 2013; Krishnan et al., 2011; McIntosh & Lobaugh, 2004; McIntosh & Mišić, 2013). PLS was introduced to the neuroimaging community as a means of finding the overlapping information across a *pattern* of brain measures and linking it to outcome variables of interest – a multivariate extension of mass univariate approaches (McIntosh et al., 1996). In turn, PLS both capitalizes on the "many to many" functional organization of the brain and is theoretically well-suited for finding robust brain mappings in the high-dimensional, multimodal datasets growing in popularity across the literature (Genon et al., 2022; Mihalik et al., 2022).

In any PLS analysis, a researcher must decide which LVs are meaningful and should be reported accordingly. Typically, this is achieved by permutation testing, which deems LVs statistically significant if they tend to explain more covariance than LVs derived from randomly shuffled data (McIntosh & Lobaugh, 2004). Importantly, during this procedure, a Procrustes rotation is applied to each set of permuted LVs to align them as closely as possible with those from the original decomposition (McIntosh & Lobaugh, 2004; Milan & Whittaker, 1995). As such, permutation testing can be thought of as comparing each latent variable with permuted latent variables capturing similar brain/behaviour associations (for more details, see *Section 2.2*).

While the Procrustes rotation is a standard component of a neuroimaging PLS workflow, the impact of this design choice on LV significance relative to the unrotated condition has not been systematically explored in the PLS literature. Moreover, the rotation is usually only applied to the smaller/behavioural component of the LVs, and it is unclear if this achieves similar results as rotating the larger/brain component, or rotating both, since PLS is fundamentally a two-sided procedure. As initial observations from our group suggested that different permutation tests (unrotated, or rotating brain/behaviour/both) performed on the same dataset can give radically different *p*-values (see *Supplementary Material, Section S1* for an overview), in the present study, we sought to determine whether each rotation method could reliably detect ground truth effects in simulated data, and whether they would lead to different conclusions regarding LV significance in samples drawn from the UK BioBank.

Finally, while LVs are generally considered robust if they pass permutation testing, there are growing concerns that the results from PLS and a range of other multivariate techniques do not necessarily reproduce in held-out samples, particularly when the analysis is performed in relatively small datasets (Marek et al., 2022; Churchill et al., 2013; Nakua et al., 2024; Ji et al., 2021; Helmer et al., 2024; Dinga et al., 2019). Similarly, some have reported that significant latent variables are not necessarily stable across samples in simulated data (Kovacevic et al., 2013; McIntosh, 2022). Building on these findings, we also evaluated whether complementary PLS outcome metrics, namely LV strength (covariance explained) and stability (across sample splits), could track ground truth effects in simulated data and identify potentially relevant LVs in the UK BioBank – above and beyond the abilities of rotated and unrotated permutation tests to do the

same. Together, then, we aimed to assess whether tests of significance should be complemented by tests of strength and stability to fully characterize a PLS decomposition.

## 2. Methods

### 2.1. The partial least squares correlation

First, we introduce how PLS is performed in the case of implementing a PLS correlation to analyze brain-behaviour covariance, sometimes referred to as behavioural PLS (Krishnan et al., 2011). Assume we have brain matrix $X$ and behavioural matrix $Y$, each with $n$ rows of observations and with z-scored feature vectors in the columns. To implement PLS, the cross-correlation matrix between $X$ and $Y$ is subjected to a singular value decomposition (SVD):

$$C = \frac{1}{n-1} X^T Y \tag{1}$$

$$C = U \Sigma V^T \tag{2}$$

With the left and right singular vectors of $C$ in the columns of $U$ and $V$, respectively, and the singular values across the diagonal of $\Sigma$. A given latent variable consists of matching columns of $U$ and $V$ (interpreted as brain and behavioural feature weights, respectively) and their corresponding singular value (proportional to covariance explained). Latent variables are mutually orthogonal and are ranked by their singular value, and the number of latent variables corresponds to the rank of $C$ (typically, the lesser of the number of columns in $X$ or $Y$). For more details, see McIntosh et al. (1996).

### 2.2. Permutation testing and the Procrustes rotation

Next, we describe how PLS permutation testing is normally implemented, with a Procrustes rotation applied to the smaller ("behavioural") component of the decomposition. First, PLS is performed as described above:

$$C = \frac{1}{n-1} X^T Y \tag{1}$$

$$C = U \Sigma V^T \tag{2}$$

Then, if we consider a single permutation instance where the rows of $Y$ are permuted, giving $Y_p$, and PLS is repeated:

$$C_p = \frac{1}{n-1} X^T Y_p \tag{3}$$

$$C_p = U_p \Sigma_p V_p^T \tag{4}$$

The orthogonal Procrustes problem (Schönemann, 1966) is solved to find the orthonormal matrix $R$ which, when applied to $V_p$, maps it as closely as possible to $V$ in a least squares sense, accounting for arbitrary changes in LV sign and order (McIntosh & Lobaugh, 2004):

$$V_p^T V = NOP^T \tag{5}$$

$$R = NP^T \tag{6}$$

Where $N$, $O$, and $P$ are the result of an SVD. For a full proof and discussion, see Schönemann (1966). Next, each row of $R$ is scaled by its corresponding diagonal element of $\Sigma_p$, and the transformation is applied to rotate (and/or reflect) the columns of $V_p$:

$$V_p' = V_p \Sigma_p R \tag{7}$$

The transformation can both reorder the singular vectors and redistribute covariance explained across them (McIntosh & Lobaugh, 2004). Then, the Euclidean norms of the scaled and transformed singular vectors are calculated:

$$val_i = ||(V_p')_i||_2 \tag{8}$$

With $i$ representing any given column of $V_p'$. The values are added to the null distributions for their corresponding latent variables (the first value for LV1, the second for LV2, etc.), and the procedure is repeated on each permutation instance. Finally, the singular values from the original decomposition, across the diagonal of $\Sigma$, are tested against their respective null distributions, giving $p$-values for each corresponding to the proportion of permuted singular values exceeding the original. Typically, LVs with $p < 0.05$ are considered statistically significant and interpreted accordingly.

### 2.3. Simulated data

First, we generated a series of datasets, each with one true latent variable present. However, our "confidence" that the effect was present varied across datasets, with more confidence corresponding to a larger sample, a stronger effect, and less noise. After performing PLS on each dataset, we could evaluate which outcome metrics (significance, strength, stability) best tracked our confidence in LV1, and by extension were the most sensitive to the dataset's key properties (see *Fig. 1* for a summary).

To create our datasets, we used the generative modeling of multivariate relationships (GEMMR) Python package, version 0.2.6 (Helmer et al., 2024). GEMMR allows users to specify the properties of a cross-correlation matrix $C$, then generate $X$ and $Y$ matrices by drawing samples from a multivariate normal distribution with mean 0 and cross-correlation matrix $C$. For more details on the simulation protocol and its assumptions, see Helmer et al. (2024).

In the first three analyses described below, we generated datasets using GEMMR and tuned a different parameter across simulations (either the sample size, canonical correlation, or noise). All unmentioned parameters were set to their default values.

#### 2.3.1. *Analysis 1: Sample size*

First, we generated 100 different datasets at each of 10 sample sizes, with sample sizes log-spaced between 10 and 10 000. All other properties of the datasets were identical, including

the number of latent effects ($m = 1$), the canonical correlation of the effect ($r = 0.3$), and the number of features in either matrix (90 $X$, 10 $Y$).

After running PLS on each dataset, we calculated the LV1 pass rate at each sample size, corresponding to the proportion of the 100 datasets with $p < 0.05$. The pass rate was our primary measure of whether a permutation test could consistently detect a simulated effect, and we directly compared pass rates between rotation methods (see *Section 2.3.5* for more details). LV1 strength and stability were also calculated at each sample size (see also *Section 2.3.5*).

#### 2.3.2. *Analysis 2: Canonical correlation*

Next, for each of 10 canonical correlations (i.e., latent effect strengths), we generated 100 different datasets, with the canonical correlations linearly spaced between 0.1 and 0.9. The number of latent effects ($m = 1$), number of features (90 $X$, 10 $Y$), and sample size ($N = 1000$) were held constant across datasets. As above, the LV1 pass rates were calculated for each permutation test method and contrasted with strength and stability metrics.

#### 2.3.3. *Analysis 3: Noise*

Finally, for simulated $X$ and $Y$ matrices, we added equally sized matrices of Gaussian noise after simulation. Specifically, we generated 100 different datasets at each of 10 noise levels, with the noise level defined by the standard deviation of a Gaussian (the 10 standard deviations were linearly spaced between 0 and 1). As above, we held all other parameters constant across datasets, including the number of latent effects ($m = 1$), the canonical correlation of the effect ($r = 0.3$), number of features (90 $X$, 10 $Y$), and sample size ($N = 1000$). LV1 pass rates (by rotation method), strength, and stability were compared accordingly.

#### 2.3.4. *Analysis 4: Whitened data*

We also wondered whether our permutation tests could "find" latent effects in datasets where the features were constrained to not meaningfully covary. Accordingly, we generated $X$ and $Y$ matrices of random values drawn from a normal distribution with mean 0 and variance 1 using the *random.randn* function from NumPy 1.26. Then, using scikit-learn 1.4.1, we standardized the columns of each matrix to have mean 0 and variance 1 with the *preprocessing.StandardScaler* function and whitened the result with the *decomposition.PCA* function. Using this procedure, we generated 100 datasets at each of 10 sample sizes, with sample sizes log-spaced between 100 and 10 000. The number of features (90 $X$, 10 $Y$) was held constant across datasets.

As above, we performed PLS on each dataset and assessed LV1 significance, strength, and stability accordingly. Importantly, given that the datasets were both randomly generated and explicitly whitened, any PLS results must be a product of sampling variability alone, rather than genuine between-feature covariance.

Note that the simulation protocol is similar to that described in *Section 2.3.1*, except the lowest sample size was set to 100 rather than 10. In turn, each matrix always had more observations than features, ensuring that it could be decorrelated.

### 2.3.5. *PLS implementation*

For a given dataset (a pair of simulated $X$ and $Y$ matrices), PLS was implemented using the *behavioral_pls* function from the pyls library in Python 3.7. We performed 4 permutation tests per dataset, each with different rotation methods, described below. Note that the "brain" and "behaviour" designation is arbitrarily based on relative matrix size.

- *None*: the rows of $Y$ were permuted and no rotation was applied to the permuted latent variables. Their singular values were directly added to the null distributions.
- *Behaviour*: as a proxy for the typically smaller behavioural matrix, the rows of $Y$ were permuted, and the Procrustes rotation was applied to its corresponding singular vectors to generate values for the null distributions (as described in *Section 2.2*).
- *Brain*: as a proxy for the typically larger brain matrix, the rows of $X$ were permuted, and the Procrustes rotation was applied to its corresponding singular vectors to generate values for the null distributions.
- *Both*: the rows of $Y$ were permuted and the Procrustes rotation was applied to both sets of singular vectors. The values returned from the two procedures were averaged together, then added to the null distributions.

10 000 permutations were performed for each permutation test, and the indices used to shuffle the matrices on a given permutation instance were identical across tests. As such, any differences in the null distributions, and by extension the $p$-values, between the permutation tests were driven by the rotations themselves rather than the reshuffling procedure.

Finally, to complement significance testing, we assessed both the covariance explained and split-half stability of the LVs. Regarding split-half stability, we employed the metric proposed by McIntosh (2022) and Nakua et al. (2024). Specifically, for a given dataset, we randomly split the sample into two halves and ran PLS on each half separately. Then, for each LV, we calculated the Pearson correlation between the singular vectors from the two halves separately for "brain" ($U$) and "behaviour" ($V$). We took the absolute value of each correlation coefficient to account for arbitrary sign flips and repeated the procedure 100 times.

## 2.4. UK BioBank

Next, we tested whether the $p$-values from the different rotation methods would disagree with each other, or diverge from strength and stability metrics, in real neuroimaging data. Accordingly, we gathered brain and behavioural data from 28 804 UK BioBank participants, with the brain matrix $X$ composed of cortical thickness in 64 regions and the behavioural matrix $Y$ composed of 17 lifestyle risk factors linked to adverse aging (see *Supplementary Materials, Section S3* for a full description of the dataset).

Then, as in the sample size analysis described in *Section 2.3.1*, we drew 100 different subsamples at each of 10 sample sizes from the broader pool of 28 804 participants, with the sample sizes log-spaced between 50 and 20 000 (the low bound ensured that split-half samples were sufficiently variable, and the high bound leveraged most of the participants while still

allowing for differences between samples). After performing PLS on each subsample, we evaluated whether significance (for each rotation method), strength, and stability depended on sample size for different LVs, as described in the previous section. The workflow is visualized in the bottom panel of *Fig. 1*. Finally, informed by our full set of simulated and UK BioBank results, we performed a PLS analysis on all 28 804 participants, and explored which measures could conceivably be used as "stopping" metrics for identifying potentially meaningful LVs.

#### 2.4.1. *PLS implementation*

For a given subsample, PLS was implemented as described in *Section 2.3.5*, with *p*-values calculated for the 4 rotation methods across 10 000 permutations. As above, covariance explained and split-half stability were also assessed for each decomposition.

Finally, for the full dataset of 28 804 participants, we implemented bootstrap resampling to assess LV feature weight reliability. Specifically, participants were sampled with replacement 10 000 times and PLS was performed on each bootstrapped sample. For each LV, we assessed bootstrap ratios for the brain features, corresponding to the ratio between a feature's mean weight and standard deviation across samples (analogous to a z-score). For the behavioural variables, we assessed loadings, or the correlation between the variable and latent variable scores across participants (with 95% confidence intervals calculated across bootstrap samples). Note that a Procrustes rotation is also typically applied to the brain component of the latent variables during bootstrap resampling (McIntosh & Lobaugh, 2004). We kept this convention for our analysis and did not assess the impact of different rotation methods on bootstrapped feature weights.

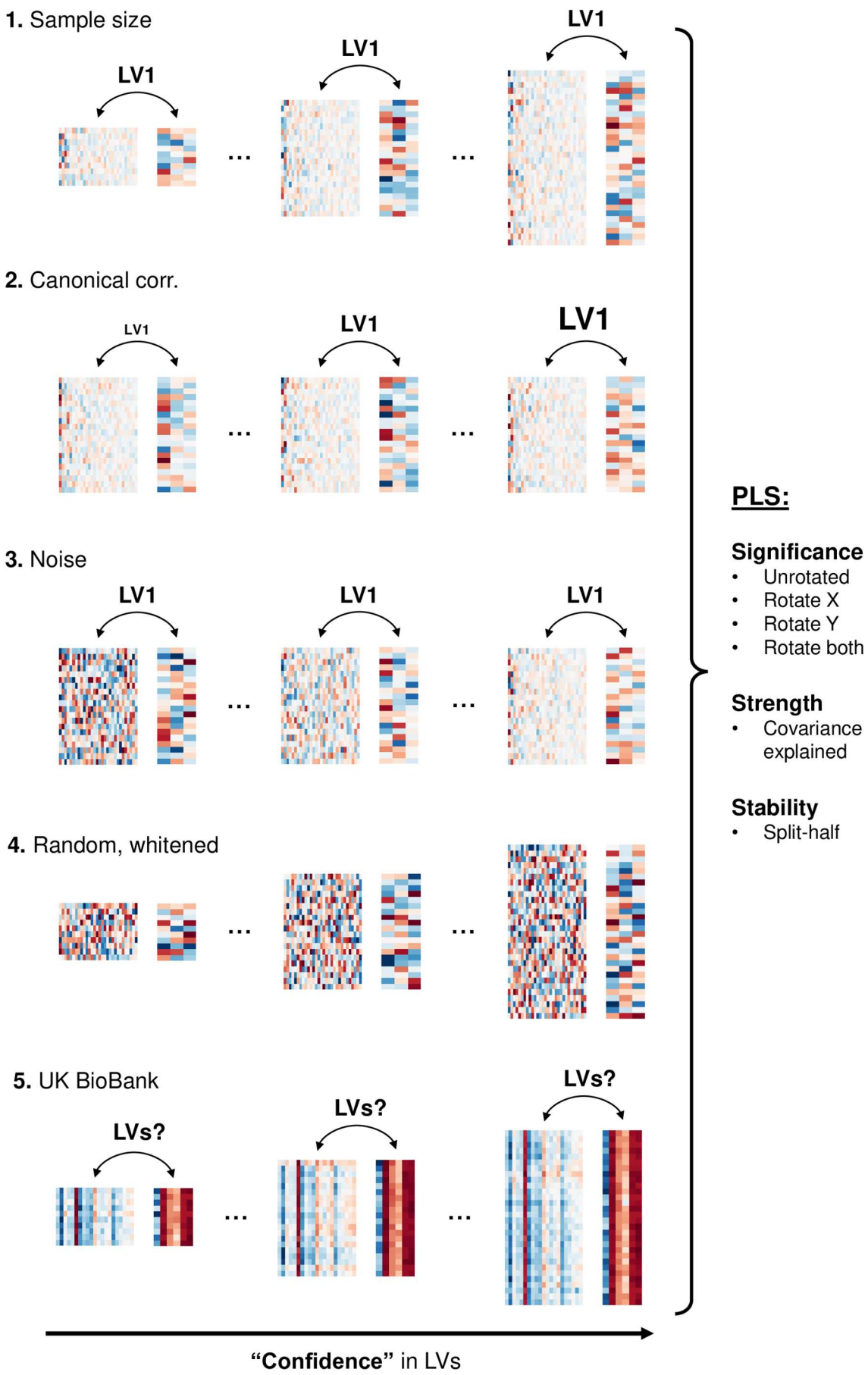
1. Sample size
LV1
LV1
LV1
2. Canonical corr.
LV1
LV1
LV1
3. Noise
LV1
LV1
LV1
4. Random, whitened
5. UK BioBank
LVs?
LVs?
LVs?
"Confidence" in LVs
PLS:
Significance
• Unrotated
• Rotate X
• Rotate Y
• Rotate both
Strength
• Covariance explained
Stability
• Split-half

**Figure 1**: workflow for the simulated and UK BioBank analyses described in *Sections 2.3-2.4*.

In the first three analyses, we generated a series of datasets with increasing "confidence" that a simulated LV1 was present, corresponding to **(1)** a larger sample, **(2)** a stronger effect, or **(3)** less noise. After performing PLS on each dataset, we expected that the listed outcome metrics would track this confidence, such that LV1 from the rightmost datasets would be stronger, more significant, and more stable than LV1 from the leftmost datasets.

Next **(4)**, we generated random, whitened datasets, expecting our PLS outcome metrics to indicate that no true effects were present (i.e., that LV1 from these datasets would be insignificant, weak, and unstable).

Finally **(5)**, we varied the size of subsamples drawn from our broader pool of UK BioBank participants, offering progressively more evidence for the "true" effects that may have been present. After performing PLS on each dataset, we evaluated whether significance, strength, and stability depended on sample size for various LVs of interest.

# 3. Results

## 3.1. Simulated data

First, we aimed to establish whether applying the Procrustes rotation during PLS permutation testing systematically impacts our abilities to detect known latent effects, and to evaluate whether complementary PLS outcome metrics, like latent variable strength and stability, can track the ground truth of a dataset more accurately than the *p*-values from any permutation test. Accordingly, we leveraged the GEMMR Python package (Helmer et al., 2024) to simulate a range of datasets, each with one true latent variable present. However, our confidence in the effect differed across datasets: either the sample size, effect strength, or amount of noise was allowed to vary, with all other parameters held constant. We then performed PLS on each dataset, and assessed which metrics (significance, strength, and stability) best represented the simulated ground truth (*Fig. 1*). Note that throughout our simulated data analyses, the larger matrix is arbitrarily referred to as "brain", while the smaller matrix is arbitrarily referred to as "behaviour", consistent with the conventions commonly used in neuroimaging PLS analyses.

### 3.1.1. *Analysis 1: Sample size*

First, we asked whether PLS would consider a latent effect to be more significant, stronger, or more stable in larger samples. Accordingly, we simulated 100 different datasets at each of 10 sample sizes, each with one latent effect present. After performing PLS on each dataset, we observed that all rotated permutation tests were more likely to detect LV1 in relatively small samples than unrotated tests, while the methods converged as *N* increased (*Fig. 2A*). Meanwhile, the covariance explained and split-half stability of LV1 both required a relatively large sample size to stabilize (*Fig. 2A*).

### *3.1.2. Analysis 2: Canonical correlation*

Next, we tested whether our PLS outcome metrics were sensitive to the strength of the simulated effect. Accordingly, we scaled the canonical correlation of our simulated latent effects, analyzing 100 different datasets at each of 10 canonical correlations. Here, all rotated permutation tests consistently detected LV1 regardless of its strength, while unrotated tests were less likely to detect relatively weak effects (*Fig. 2B*). Predictably, covariance explained climbed with the simulated canonical correlation, and split-half stability did not stabilize until the effect was relatively strong (*Fig. 2B*).

### 3.1.3. *Analysis 3: Noise*

Next, we assessed whether noise could compromise the sensitivity of PLS to ground truth effects. Analyzing 100 different datasets at each of 10 noise levels, we observed that all rotated permutation tests consistently detected LV1 regardless of the amount of noise, while the unrotated tests were more conservative in noisier datasets (*Fig. 2C*). Alongside unrotated significance, the covariance explained and split-half stability of LV1 both decreased as noise levels increased (*Fig. 2C*).

### *3.1.4. Analysis 4: Whitened data*

So far, our results have suggested that rotated permutation tests tend to detect LV1 under most conditions. As such, we asked whether LV1 could still be considered significant when PLS was performed on randomly generated and explicitly whitened data – in other words, data with no meaningful between-feature covariance. Varying the sample size of the whitened datasets, we found that all rotated permutation tests considered LV1 significant regardless of *N*, while the unrotated tests rarely detected the unsimulated "effect" (*Fig. 2D*). Meanwhile, LV1 strength and stability metrics increased marginally with sample size, though both sets of metrics plateaued at relatively low values (*Fig. 2D*).

### 3.1.5. *Summary*

Together, we observed that rotated permutation tests consistently detected LV1, regardless of whether the effect was present in a small sample, weak, obscured by noise, or present at all. This was true regardless of which LV component was rotated ("behaviour", "brain", or both). Notably, rotated tests also consistently detected LV2, which was not simulated in any analysis (*Supplementary Fig. S5*). Meanwhile, unrotated significance, covariance explained, and split-half stability accurately tracked our confidence that a simulated effect was present, with more confidence corresponding to a larger sample, a stronger canonical correlation, and less noise.

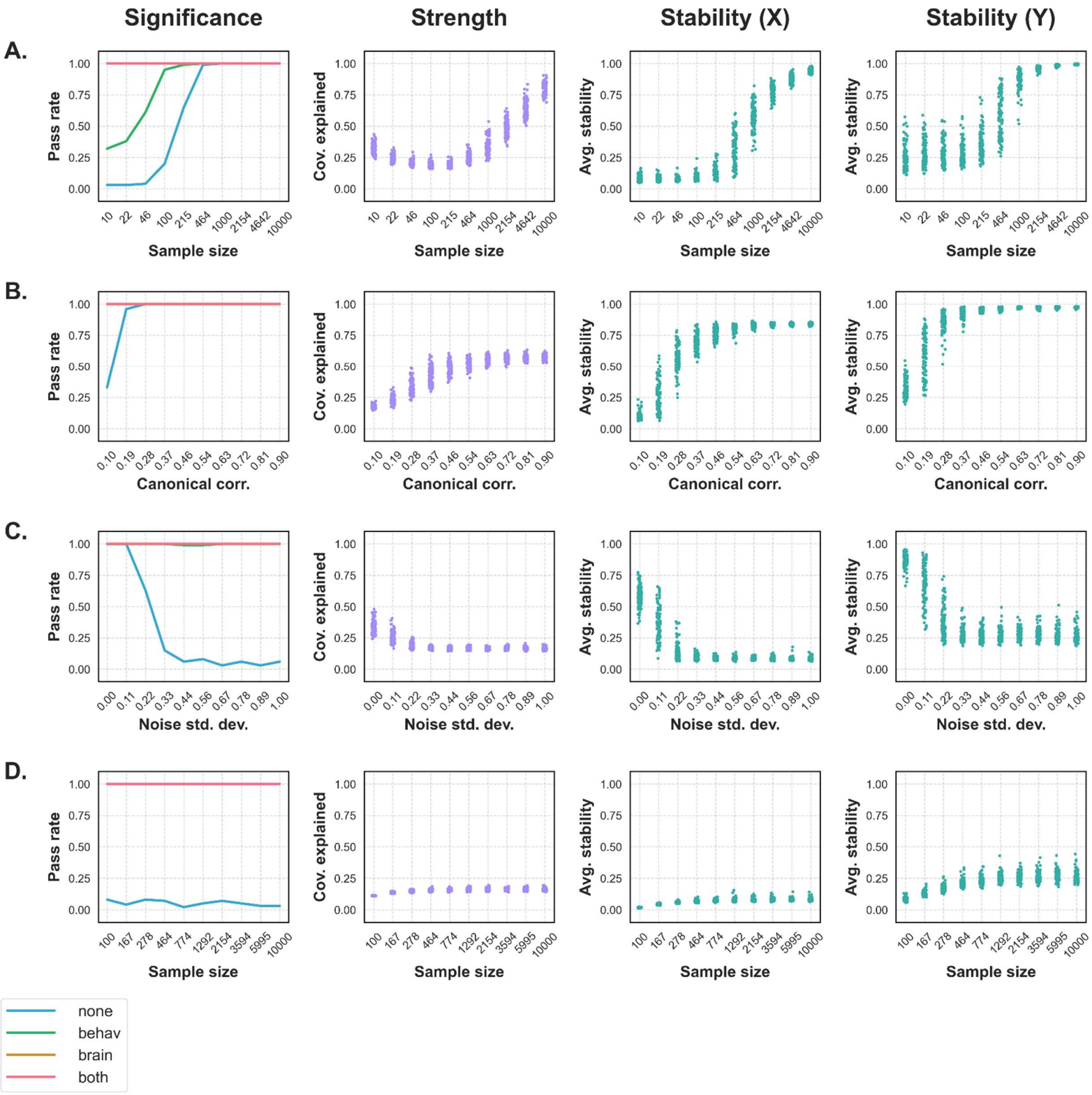
Significance
Strength
Stability (X)
Stability (Y)
A.
B.
C.
D.
Pass rate
Cov. explained
Avg. stability
Sample size
Canonical corr.
Noise std. dev.
none
behav
brain
both

**Figure 2**: LV1 significance, strength, and stability metrics across each simulated data analysis.

The **Significance** column shows the pass rates for LV1, or the proportion of the 100 datasets with a significant ($p < 0.05$) LV1 at each step, plotted separately for each rotation method. The **Strength** column shows the covariance explained of LV1, and the **Stability (X)** and **Stability (Y)** columns show the average split-half stability of LV1 across 100 splits for the "brain" and "behaviour" LV components, respectively. Individual points represent values for unique simulated datasets.

**A. (Sample size)** Pass rates steadily increased with *N* when no rotation was applied, while all rotated permutation tests were more likely to detect the effect in smaller samples. Like the unrotated pass rates, LV1 strength and stability both climbed with sample size.

**B. (Canonical correlation)** All rotated permutation tests consistently detected the first latent variable regardless of how strong it was, while the unrotated test was less sensitive to the weakest effects. Meanwhile, strength and stability both increased with the simulated canonical correlation.

**C. (Noise)** The unrotated permutation test failed to consistently detect the simulated effect under noisy conditions, while all rotated tests generally considered LV1 significant regardless of noise levels. Alongside unrotated significance, the strength and stability of LV1 all decreased as noise levels increased.

**D. (Whitened data)** Rotated permutation tests consistently deemed LV1 significant, even though no effects were encoded. Meanwhile, the unrotated test was far less likely to find the unsimulated "effect", and strength and stability metrics remained low regardless of sample size.

### 3.2. UK BioBank

In our simulated data analysis, we saw that latent variable significance differed between the rotated and unrotated conditions, and that significant latent variables were not necessarily strong or stable if a rotation was applied during permutation testing. Accordingly, we wondered whether our PLS outcome metrics would also offer non-overlapping information in a real neuroimaging dataset. Specifically, we analyzed brain (regional cortical thickness, $X$) and behavioural (lifestyle risk factors linked to adverse aging, $Y$) data from 28 804 UK BioBank participants (for more details, see *Supplementary Materials – Section S3*). As in our simulated data analysis, we performed PLS on subsamples drawn from the broader pool of participants, asking if the significance, strength, and stability of different LVs depended on sample size. Informed by our results, we performed a final PLS analysis on the full set of 28 804 participants, and explored which outcome metrics could potentially discriminate meaningful LVs.

#### 3.2.1. *Analysis 1: Sample size*

Drawing 100 random subsamples at each of 10 sample sizes from our pool of UK BioBank participants, we saw that the number of significant brain-behaviour latent variables increased with $N$ (*Fig. 3*). In this regard, the unrotated test was the most liberal method (passing nearly every LV of a possible 17 when $N$ was large, on average), while rotating the behavioural component was the most conservative. While unrotated tests tended to consider late LVs significant in large samples, the split-half stability and covariance explained of these effects tended to be relatively low. Sample plots showing these trends across LVs 1, 3, 5, and 10 are shown in *Supplementary Fig. S7*.

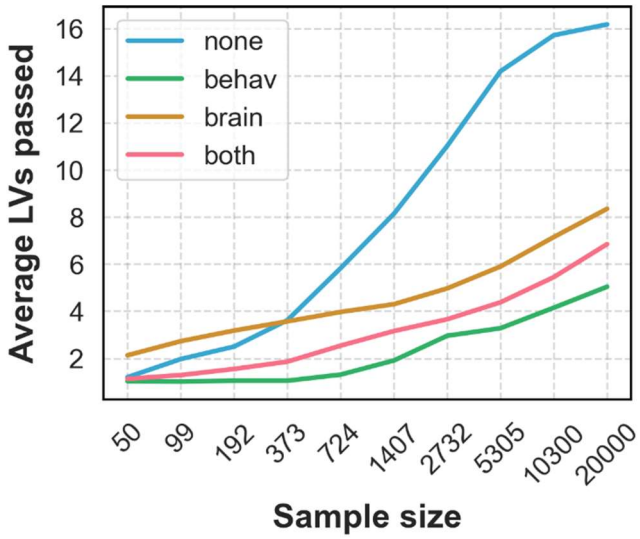


**Figure 3**: the average number of latent variables passing each permutation test ($p < 0.05$) across the 100 subsamples at each sample size. While the number of significant LVs increased with $N$ for each method, the effect was especially pronounced for the unrotated tests, which passed nearly every possible LV in large samples.

### 3.2.2. *Analysis 2: Full sample*

In our full sample of 28 804 participants, all permutation tests agreed that the first 5 LVs were significant (*Fig. 4*). However, the methods disagreed for further latent variables, with the unrotated test determining that *all* possible LVs were significant.

While the unrotated test was entirely non-selective, our previous results suggested that rotated tests are systematically permissive of early LVs. Owing to the limitations of both approaches, we instead turn to latent variable strength and stability metrics, which tracked key properties of our simulated data, to determine which effects may be of interest in our full UK BioBank sample.

For one, split-half stability became far more variable across splits after LV4 (*Fig. 4*). In this case, a researcher may choose to set a stability threshold and report the first 4 LVs accordingly. Beyond being relatively stable, the first 4 LVs were also relatively interpretable. LV1 reflected globally thicker cortex in younger participants, while LV2, LV3, and LV4 reflected more focal cortical thickness patterns related most strongly to alcohol consumption, body-mass index, and education/smoking, respectively (*Fig. 5*). Behavioural loadings for subsequent LVs tended to be relatively high for variables which already loaded strongly onto previous LVs, or tended to be relatively scattered (*Supplementary Fig. S8*). In turn, a researcher may choose to report the first 4 latent variables, owing to their relative stability *and* coherence.

As a final example, a researcher may instead choose to only report the first latent variable, as it explained most of the covariance in the sample (*Fig. 4*). This may especially be the case if their research question is age-related, as LV1 is the dominant age-related pattern in the dataset, and all subsequent latent variables must capture covariance independent of it (see *Section 2.1*).

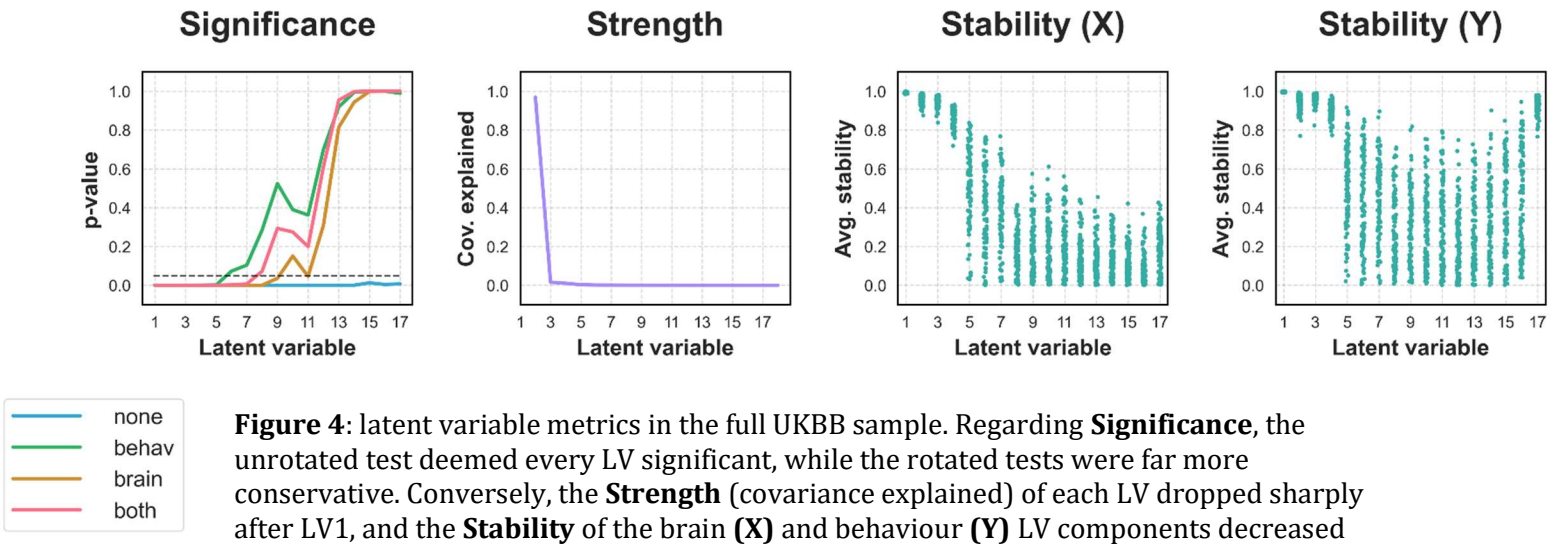


**Figure 4**: latent variable metrics in the full UKBB sample. Regarding **Significance**, the unrotated test deemed every LV significant, while the rotated tests were far more conservative. Conversely, the **Strength** (covariance explained) of each LV dropped sharply after LV1, and the **Stability** of the brain **(X)** and behaviour **(Y)** LV components decreased and became far more variable after LV4 (individual points represent values for unique dataset splits).

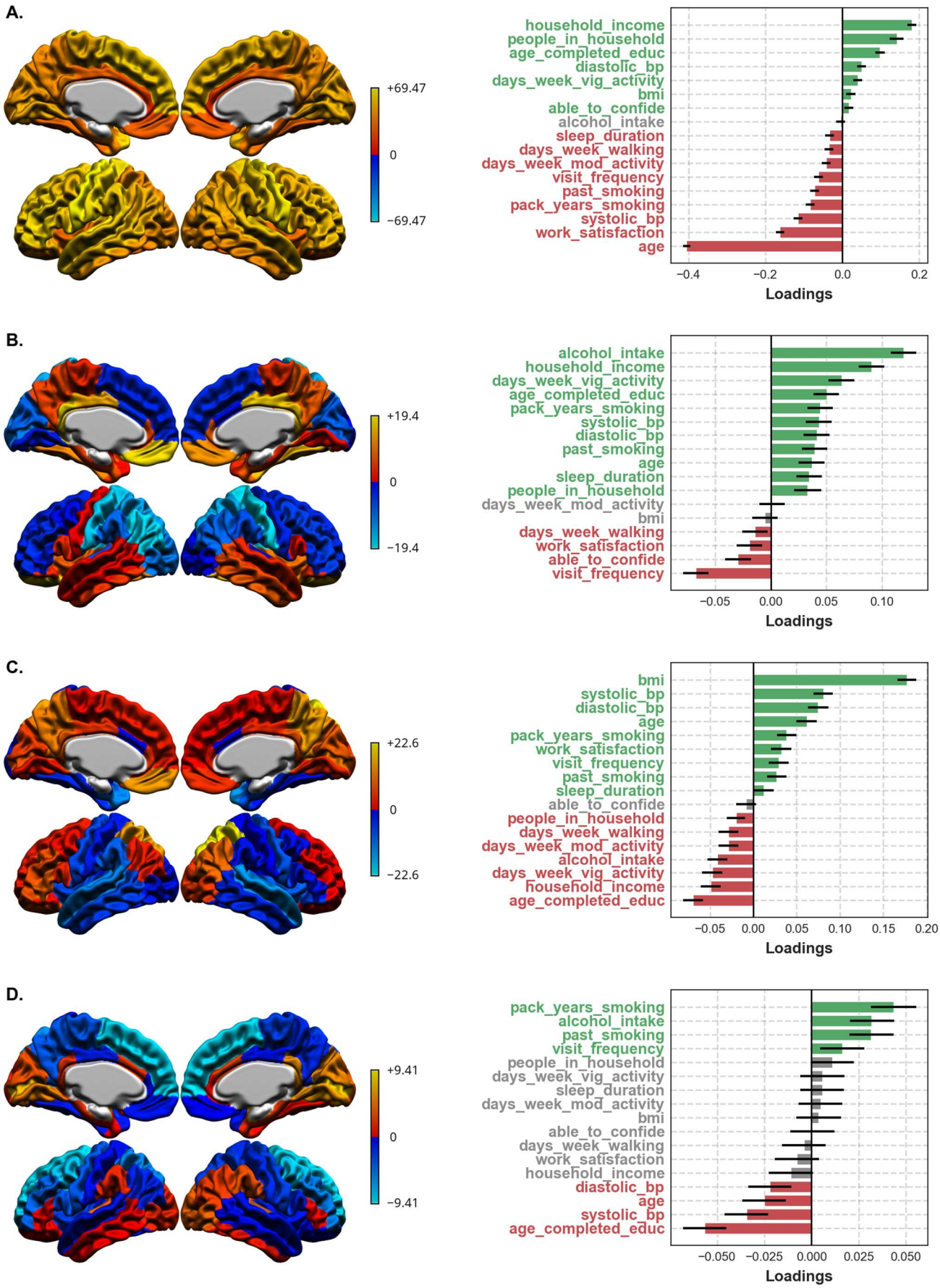
A.
+69.47
0
−69.47
household_income
people_in_household
age_completed_educ
diastolic_bp
days_week_vig_activity
bmi
able_to_confide
alcohol_intake
sleep_duration
days_week_walking
days_week_mod_activity
visit_frequency
past_smoking
pack_years_smoking
systolic_bp
work_satisfaction
age
−0.4
−0.2
0.0
0.2
Loadings
B.
+19.4
0
−19.4
alcohol_intake
household_income
days_week_vig_activity
age_completed_educ
pack_years_smoking
systolic_bp
diastolic_bp
past_smoking
age
sleep_duration
people_in_household
days_week_mod_activity
bmi
days_week_walking
work_satisfaction
able_to_confide
visit_frequency
−0.05
0.00
0.05
0.10
Loadings
C.
+22.6
0
−22.6
bmi
systolic_bp
diastolic_bp
age
pack_years_smoking
work_satisfaction
visit_frequency
past_smoking
sleep_duration
able_to_confide
people_in_household
days_week_walking
days_week_mod_activity
alcohol_intake
days_week_vig_activity
household_income
age_completed_educ
−0.05
0.00
0.05
0.10
0.15
0.20
Loadings
D.
+9.41
0
−9.41
pack_years_smoking
alcohol_intake
past_smoking
visit_frequency
people_in_household
days_week_vig_activity
sleep_duration
days_week_mod_activity
bmi
able_to_confide
days_week_walking
work_satisfaction
household_income
diastolic_bp
age
systolic_bp
age_completed_educ
−0.050
−0.025
0.000
0.025
0.050
Loadings

**Figure 5**: LVs 1-4 in the full UKBB sample. The first column shows bootstrap ratios for the brain variables, while the second shows loadings for the behavioural variables.

**A. (LV1)** An expected pattern of broadly thicker cortex in younger participants.

**B. (LV2)** A pattern of increases and decreases in cortical thickness across frontal, parietal, and cingulate regions related to greater alcohol consumption.

**C. (LV3)** Broadly thicker cortex in the frontal and parietal lobes related to greater body-mass index.

**D. (LV4)** Thinner superior frontal cortex related to greater smoking and lower education.

## 4. Discussion

In a PLS analysis, latent variables are considered statistically significant if they pass permutation testing, with a Procrustes rotation applied to realign each set of permuted latent variables. However, it has not been shown whether applying the rotation, or applying the rotation to different components of the decomposition, provides a more accurate measure of whether true latent variables are present in a dataset, relative to the unrotated condition. Further, given growing concerns that the results from PLS and other multivariate techniques can fail to reproduce in held-out samples (Marek et al., 2022; Churchill et al., 2013; Nakua et al., 2024; Ji et al., 2021; Helmer et al., 2024; Dinga et al., 2019), it is plausible that significance testing can be complemented by tests based on latent variable strength and stability to more thoroughly evaluate a PLS decomposition.

Accordingly, across a range of simulated datasets with a known ground truth, we observed that rotated permutation tests routinely detected the first latent variable, regardless of the rotation method applied and regardless of whether the effect was undersampled, weak, noisy, or drawn from data with no meaningful between-feature covariance. Unrotated tests were more conservative in simulated data, but proved to be non-selective in large samples drawn from the UK BioBank. Despite the limitations of the permutation tests, strength and stability metrics were relatively coherent: they accurately tracked our confidence that latent effects were present in simulated data, and showed potential for discriminating effects of interest in the UK BioBank.

The tendency for rotated tests to favour passing LV1 agrees with previous observations (Kovacevic et al., 2013) and follows from our understanding of PLS. Critically, in any PLS decomposition, LV1 is the *strongest possible* source of covariance in the sample. Therefore, after PLS is performed on a given permuted dataset and a Procrustes rotation is applied to redistribute covariance across the permuted LVs (see *Section 2.2 – Equation 7*), there is no situation where the permuted LV1 can "gain" covariance, as it was already as strong as possible before the rotation. In turn, following a rotation, covariance must tend to "shift" from early to late null distributions – making it easier to detect the strongest effect, and more difficult to detect weaker ones. For visual examples of this effect in whitened and UK BioBank data, see *Supplementary Figs. S6* and *S9*. Finally, the tendency for the Procrustes rotation to "dilute" the first of a set of magnitude-ordered effects should extend to other applications of the technique beyond PLS permutation testing, including recent gradient mapping studies, where the rotation is used to align axes describing continuous cortical variability in an individual to those from a group average (Langs et al., 2015; Vos de Wael et al., 2020).

More generally, it is unclear whether applying the rotation during permutation testing has a strong theoretical backing. For one, the covariance explained by LV1 in any sample should overestimate that of the population (Lawley, 1956). If the strength of LV1 is already overestimated, and if rotated permutation tests make LV1 easier to detect, then the procedure may be suboptimal for PLS statistical inference. Further, the Procrustes rotation was originally brought to PLS and related techniques in the context of bootstrap resampling, not permutation testing (Milan & Whittaker, 1995). Relative to the original sample, any bootstrapped sample

should contain the same latent effects, though they may be reordered due to sampling variability – justifying a transformation to map them to the originals (Milan & Whittaker, 1995). However, it is unclear whether such a situation holds for the permutation test, where latent effects are not simply reordered, but fundamentally *broken* by the reshuffling procedure. Finally, it should be noted that there is active research in the broader field of Procrustes analysis, which studies the optimal means of mapping one matrix to another. The orthogonal Procrustes problem, solved in PLS permutation testing, is just one way to perform a Procrustes analysis, and several other methods could equally be employed (a detailed discussion is beyond the scope of this report – see Gower (2010), Meng et al. (2022) for more details).

Beyond the permutation testing approach considered here, an alternative technique was recently proposed by Nakua et al. (2024). Briefly, the test considers whether the total covariance explained by a *set* of latent variables is greater than expected by chance (i.e.., the sum of the covariance captured by LV2-LV*N* must be significant for LV2 to be considered significant). This approach is similar to permutation inference based on the Wilk's λ statistic, commonly used for canonical correlation analysis, a sister technique of PLS (Winkler et al., 2020). Importantly, the approach both avoids applying a Procrustes rotation and is thought to perform well in large samples (Nakua et al., 2024), obviating the concerns raised in this report about unrotated tests in large samples and rotated tests more generally.

Alongside permutation testing, we also assessed whether latent variable strength and stability were sensitive to key properties of simulated data. Tellingly, the split-half stability and covariance explained of LV1 tended to be relatively low when the simulated effect was undersampled, weak, or noisy, and when PLS was performed on whitened data. Our findings agree with previous PLS simulation studies, which have reported that stability metrics were sensitive to sample size (Helmer et al., 2024), effect strength (Kovacevic et al., 2013), noise (Kovacevic et al., 2013), and whether an effect was encoded (McIntosh, 2022). Accordingly, beyond PLS, our results add to a growing body of literature showing that significant effects are not necessarily strong or stable across samples, stressing the importance of incorporating statistical analyses beyond significance testing, especially in multivariate paradigms (Chen et al., 2017; Dinga et al., 2019; Marek et al., 2022).

While we demonstrated that latent variable strength and stability metrics were coupled to a simulated ground truth, we did not provide a framework based on these measures for selecting the "true" latent variables in a dataset. However, many such frameworks have been described across the PLS literature. For one, Carbonell (2020) simply reported LV1, which by definition captured the most covariance in the sample (regardless, one could argue that any effects after LV1 are somewhat artificial, as they are all constrained to be independent of it). Meanwhile, Misic et al. (2016) analyzed LVs which passed two tests for covariance explained commonly seen in the principal components analysis literature. Specifically, they tested whether an LV was stronger than the average LV in the dataset, and whether it passed a visual scree test for covariance explained. Regarding split-half stability, McIntosh (2022) and Nakua et al. (2024) calculated *z*-values describing the distributions of stability values for each LV, and suggested that LVs with a

*z*-value above a heuristic threshold can be considered consistently stable. Finally, much like assessing stability, Hansen et al. (2021), Kirschner et al. (2020), and Mirchi et al. (2019) all cross-validated their PLS results by assessing whether the cross-modality relationships captured by LVs in a training dataset were sufficiently strong in a held-out sample. Based on our findings, all such approaches may serve as reasonable alternatives or complements to conventional permutation testing.

Together, rotated permutation tests are systematically more permissive of early latent variables, while unrotated permutation tests can be exceedingly liberal in large samples. Owing to the limitations of both approaches, we argue that the *p*-values from either test cannot be interpreted as strict measures of whether a latent variable "exists", and that tests based on latent variable strength and stability should be considered as part of a more nuanced formula for determining the effects to report in a PLS analysis. We end by presenting a set of considerations based on our findings for researchers implementing PLS permutation testing, shown in *Table 1*.

| | **Considerations** |
|---|---|
| *1. Significance testing* | Significant latent variables are not necessarily meaningful, regardless of the rotation applied during permutation testing. |
| *2. Rotated permutation tests* | Rotated permutation tests tend to systematically pass early latent variables. Complementary tests based on latent variable strength and stability can help determine whether an effect may be meaningful. |
| *3. Unrotated permutation tests* | Unrotated permutation tests tend to be overly permissive when *N* is large. More conservative methods are preferred in large datasets. |
| *4. Procrustes analysis* | Research is ongoing into how to best align one matrix with another. The rotation applied during PLS permutation testing is not necessarily optimal. |

***Table 1****:* considerations for PLS permutation testing.

## References

Abdi, H., Chin, W. W., Esposito Vinzi, V., Russolillo, G., & Trinchera, L. (Eds.). (2013). *New Perspectives in Partial Least Squares and Related Methods* (Vol. 56). Springer New York. https://doi.org/10.1007/978-1-4614-8283-3

Benoit, A., Malla, A. K., Iyer, S. N., Joober, R., Bherer, L., & Lepage, M. (2015). Cognitive deficits characterization using the CogState Research Battery in first-episode psychosis patients. *Schizophrenia Research. Cognition*, *2*(3), 140–145. https://doi.org/10.1016/j.scog.2015.03.006

Carbonell, F., Zijdenbos, A. P., Bedell, B. J., & Alzheimer's Disease Neuroimaging Initiative. (2020). Spatially Distributed Amyloid-β Reduces Glucose Metabolism in Mild Cognitive Impairment. *Journal of Alzheimer's Disease: JAD*, *73*(2), 543–557. https://doi.org/10.3233/JAD-190560

Chen, G., Taylor, P. A., & Cox, R. W. (2017). Is the statistic value all we should care about in neuroimaging? *NeuroImage*, *147*, 952–959. https://doi.org/10.1016/j.neuroimage.2016.09.066

Churchill, N., Spring, R., Abdi, H., Kovacevic, N., McIntosh, A. R., & Strother, S. (2013). The Stability of Behavioral PLS Results in Ill-Posed Neuroimaging Problems. In H. Abdi, W. W. Chin, V. Esposito Vinzi, G. Russolillo, & L. Trinchera (Eds.), *New Perspectives in Partial Least Squares and Related Methods* (pp. 171–183). Springer. https://doi.org/10.1007/978-1-4614-8283-3_11

Dinga, R., Schmaal, L., Penninx, B. W. J. H., van Tol, M. J., Veltman, D. J., van Velzen, L., Mennes, M., van der Wee, N. J. A., & Marquand, A. F. (2019). Evaluating the evidence for biotypes of depression: Methodological replication and extension of. *NeuroImage. Clinical*, *22*, 101796. https://doi.org/10.1016/j.nicl.2019.101796

Genon, S., Eickhoff, S. B., & Kharabian, S. (2022). Linking interindividual variability in brain structure to behaviour. *Nature Reviews Neuroscience*, *23*(5), Article 5. https://doi.org/10.1038/s41583-022-00584-7

Gower, J. C. (2010). Procrustes methods. *WIREs Computational Statistics*, *2*(4), 503–508. https://doi.org/10.1002/wics.107

Hansen, J. Y., Markello, R. D., Vogel, J. W., Seidlitz, J., Bzdok, D., & Misic, B. (2021). Mapping gene transcription and neurocognition across human neocortex. *Nature Human Behaviour*, *5*(9), 1240–1250. https://doi.org/10.1038/s41562-021-01082-z

Helmer, M., Warrington, S., Mohammadi-Nejad, A.-R., Ji, J. L., Howell, A., Rosand, B., Anticevic, A., Sotiropoulos, S. N., & Murray, J. D. (2024). On the stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. *Communications Biology*, *7*(1), 1–15. https://doi.org/10.1038/s42003-024-05869-4

Ji, J. L., Helmer, M., Fonteneau, C., Burt, J. B., Tamayo, Z., Demšar, J., Adkinson, B. D., Savić, A., Preller, K. H., Moujaes, F., Vollenweider, F. X., Martin, W. J., Repovš, G., Cho, Y. T., Pittenger, C., Murray, J. D., & Anticevic, A. (2021). Mapping brain-behavior space relationships along the psychosis spectrum. *eLife*, *10*, e66968. https://doi.org/10.7554/eLife.66968

Kim, J. S., Singh, V., Lee, J. K., Lerch, J., Ad-Dab'bagh, Y., MacDonald, D., Lee, J. M., Kim, S. I., & Evans, A. C. (2005). Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. *NeuroImage*, *27*(1), 210–221. https://doi.org/10.1016/j.neuroimage.2005.03.036

Kirschner, M., Shafiei, G., Markello, R. D., Makowski, C., Talpalaru, A., Hodzic-Santor, B., Devenyi, G. A., Paquola, C., Bernhardt, B. C., Lepage, M., Chakravarty, M. M., Dagher, A., & Mišić, B. (2020). Latent Clinical-Anatomical Dimensions of Schizophrenia. *Schizophrenia Bulletin*, *46*(6), 1426–1438. https://doi.org/10.1093/schbul/sbaa097

Kovacevic, N., Abdi, H., Beaton, D., & McIntosh, A. (2013). Revisiting PLS Resampling: Comparing Significance Versus Reliability Across Range of Simulations. *Springer Proceedings in Mathematics and Statistics*, *56*, 159–170. https://doi.org/10.1007/978-1-4614-8283-3_10

Krishnan, A., Williams, L. J., McIntosh, A. R., & Abdi, H. (2011). Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage*, *56*(2), 455–475. https://doi.org/10.1016/j.neuroimage.2010.07.034

Langs, G., Golland, P., & Ghosh, S. S. (2015). Predicting Activation Across Individuals with Resting-State Functional Connectivity Based Multi-Atlas Label Fusion. *Medical Image Computing and Computer-Assisted Intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, *9350*, 313–320. https://doi.org/10.1007/978-3-319-24571-3_38

Lawley, D. N. (1956). Tests of Significance for the Latent Roots of Covariance and Correlation Matrices. *Biometrika*, *43*(1/2), 128–136. https://doi.org/10.2307/2333586

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), Article 7902. https://doi.org/10.1038/s41586-022-04492-9

McIntosh, A. R. (2022). *Comparison of Canonical Correlation and Partial Least Squares analyses of simulated and empirical data* (arXiv:2107.06867). arXiv. https://doi.org/10.48550/arXiv.2107.06867

McIntosh, A. R., Bookstein, F. L., Haxby, J. V., & Grady, C. L. (1996). Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage*, *3*(3 Pt 1), 143–157. https://doi.org/10.1006/nimg.1996.0016

McIntosh, A. R., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: Applications and advances. *NeuroImage*, *23 Suppl 1*, S250-263. https://doi.org/10.1016/j.neuroimage.2004.07.020

McIntosh, A. R., & Mišić, B. (2013). Multivariate Statistical Analyses for Neuroimaging Data. *Annual Review of Psychology*, *64*(1), 499–525. https://doi.org/10.1146/annurev-psych-113011-143804

Meng, F., Richer, M., Tehrani, A., La, J., Kim, T. D., Ayers, P. W., & Heidar-Zadeh, F. (2022). Procrustes: A python library to find transformations that maximize the similarity between matrices. *Computer Physics Communications*, *276*, 108334. https://doi.org/10.1016/j.cpc.2022.108334

Mihalik, A., Chapman, J., Adams, R. A., Winter, N. R., Ferreira, F. S., Shawe-Taylor, J., & Mourão-Miranda, J. (2022). Canonical Correlation Analysis and Partial Least Squares for Identifying Brain–Behavior Associations: A Tutorial and a Comparative Study. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *7*(11), 1055–1067. https://doi.org/10.1016/j.bpsc.2022.07.012

Milan, L., & Whittaker, J. (1995). Application of the Parametric Bootstrap to Models that Incorporate a Singular Value Decomposition. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *44*(1), 31–49. https://doi.org/10.2307/2986193

Mirchi, N., Betzel, R. F., Bernhardt, B. C., Dagher, A., & Mišić, B. (2019). Tracking mood fluctuations with functional network patterns. *Social Cognitive and Affective Neuroscience*, *14*(1), 47–57. https://doi.org/10.1093/scan/nsy107

Mišić, B., Betzel, R. F., de Reus, M. A., van den Heuvel, M. P., Berman, M. G., McIntosh, A. R., & Sporns, O. (2016). Network-Level Structure-Function Relationships in Human Neocortex. *Cerebral Cortex*, *26*(7), 3285–3296. https://doi.org/10.1093/cercor/bhw089

Nakua, H., Yu, J.-C., Abdi, H., Hawco, C., Voineskos, A., Hill, S., Lai, M.-C., Wheeler, A. L., McIntosh, A. R., & Ameis, S. H. (2024). Comparing the stability and reproducibility of brain-behaviour relationships found using Canonical Correlation Analysis and Partial Least Squares within the ABCD Sample. *Network Neuroscience*, 1–52. https://doi.org/10.1162/netn_a_00363

Pruessner, M., Faridi, K., Shah, J., Rabinovitch, M., Iyer, S., Abadi, S., Pawliuk, N., Joober, R., & Malla, A. K. (2017). The Clinic for Assessment of Youth at Risk (CAYR): 10 years of service delivery and research targeting the prevention of psychosis in Montreal, Canada. *Early Intervention in Psychiatry*, *11*(2), 177–184. https://doi.org/10.1111/eip.12300

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, *31*(1), 1–10. https://doi.org/10.1007/BF02289451

Vos de Wael, R., Benkarim, O., Paquola, C., Lariviere, S., Royer, J., Tavakol, S., Xu, T., Hong, S.-J., Langs, G., Valk, S., Misic, B., Milham, M., Margulies, D., Smallwood, J., & Bernhardt, B. C. (2020). BrainSpace: A toolbox for the analysis of macroscale gradients in neuroimaging and connectomics datasets. *Communications Biology*, *3*(1), 1–10. https://doi.org/10.1038/s42003-020-0794-7

Winkler, A. M., Renaud, O., Smith, S. M., & Nichols, T. E. (2020). Permutation inference for canonical correlation analysis. *NeuroImage*, *220*, 117065. https://doi.org/10.1016/j.neuroimage.2020.117065

## Data and Code Availability

The code used to reproducibly generate the simulated datasets and perform the data analysis is available on GitHub.

## Author Contributions

**MD:** Conceptualization, Software, Formal Analysis, Investigation, Writing – Original Draft, Writing – Review & Editing. **YZ:** Conceptualization, Formal Analysis, Writing – Review & Editing. **AM:** Data Curation. **ML:** Data Curation. **JS:** Data Curation. **RJ:** Data Curation. **BM:** Conceptualization, Writing – Review & Editing. **YIM:** Conceptualization, Supervision, Writing – Review & Editing. **MMC:** Conceptualization, Supervision, Writing – Review & Editing, Funding Acquisition.

## Funding

This work was undertaken thanks in part to support from the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), le Fonds de recherche du Québec – Santé (FRQS), and the Canada First Research Excellence Fund (CFREF), awarded through the Healthy Brains, Healthy Lives (HBHL) initiative at McGill University.

## Declaration of Competing Interests

The authors have no competing interests to declare.

## Acknowledgements

The authors thank T. Robert Baumeister for helpful initial conversations.

# S. Supplementary Material

## S1. Preliminary findings

Data were collected through the recruitment of clients and their families at the *Prevention and Early Intervention for Psychosis* and *Clinic for Assessment of Youth at Risk* services at the Douglas Mental Health University Institute in Montreal, Canada (Pruessner et al., 2017). Participants were either diagnosed with a first episode of psychosis within 3 months of study participation, at clinical high-risk for psychosis (expressing subthreshold symptoms according to CAARMS criteria (Pruessner et al., 2017)), or at familial high-risk for psychosis (siblings of the psychosis patients in the study). All participants provided written informed consent for the secondary analysis of their data, and the study was approved by a Research Ethics Board at the Douglas Mental Health University Institute.

All images were recorded with a 3T Simens Magnetom scanner. T1-weighted images (MPRAGE, 1 $mm^3$) were collected for each participant and preprocessed using the minc-bpipe-library. Images were then processed with CIVET 2.1.1 (Kim et al., 2005), giving cortical thickness at 77 122 vertices across the cortical surface. Positive and negative psychotic symptoms were measured using the 9 global items from the interviewer-assessed Scales for the Assessment of Positive/Negative Symptoms, and cognitive performance was summarized across the 7 domains of the CogState computerized research battery (Benoit et al., 2015).

Our final sample included 125 participants whose images passed strict motion and CIVET quality control guidelines and whose symptom measures were complete. Accordingly, for our PLS analysis, the brain matrix was composed of (125 · 77 122) vertex-wise cortical thickness measures and the behavioural matrix was composed of (125 · 16) clinical and cognitive measures. PLS was performed as described in *Section 2.3.5* with the 4 permutation test methods. To confirm that our results were not driven by an inadequate number of permutations, the input matrices were reshuffled 100 000 times.

The *p*-values by LV for each permutation test are shown in *Supplementary Fig. S1.* All tests agreed that LV1 was significant, but the *p*-values for LV2 dramatically diverged, ranging from significant when rotating the brain component to approaching 1 when no rotation was applied. For reference, the LV2 null distributions are shown in *Supplementary Fig. S2*.

As LV2 was relatively coherent, linking focal cortical thickness reductions to a clear pattern of positive and disorganized psychotic symptoms (*Supplementary Fig. S3*), we sought to systematically evaluate which permutation test should be "trusted" regarding its significance, motivating the present study.

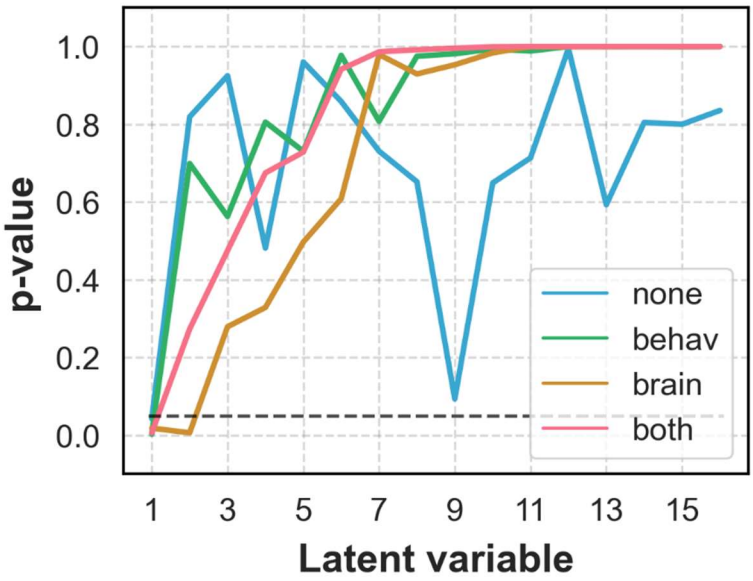


**Figure S1**: *p*-values by latent variable for each permutation test performed in our psychosis spectrum dataset. The dashed line indicates a conventional significance threshold ($p < 0.05$). *p*-values began to meaningfully diverge after LV1.

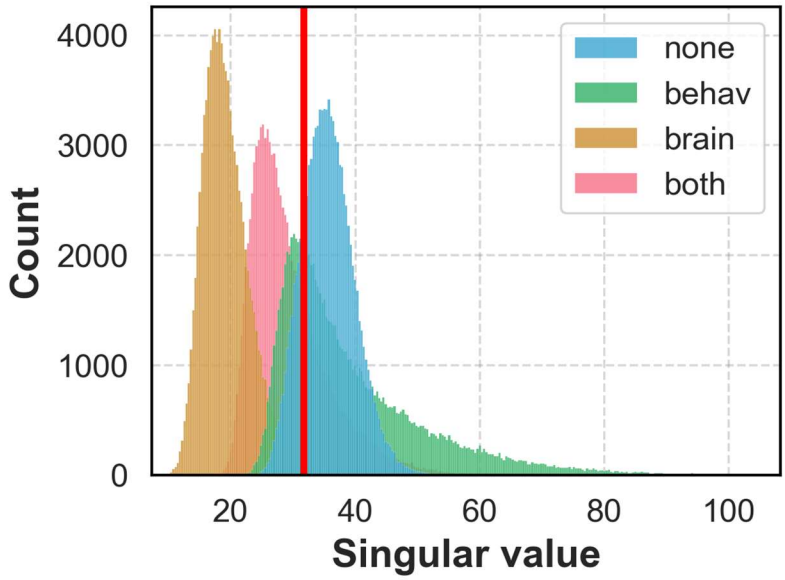


**Figure S2:** LV2 null distributions for each permutation test performed in our psychosis spectrum dataset. The red line represents the LV2 singular value from the initial PLS decomposition.

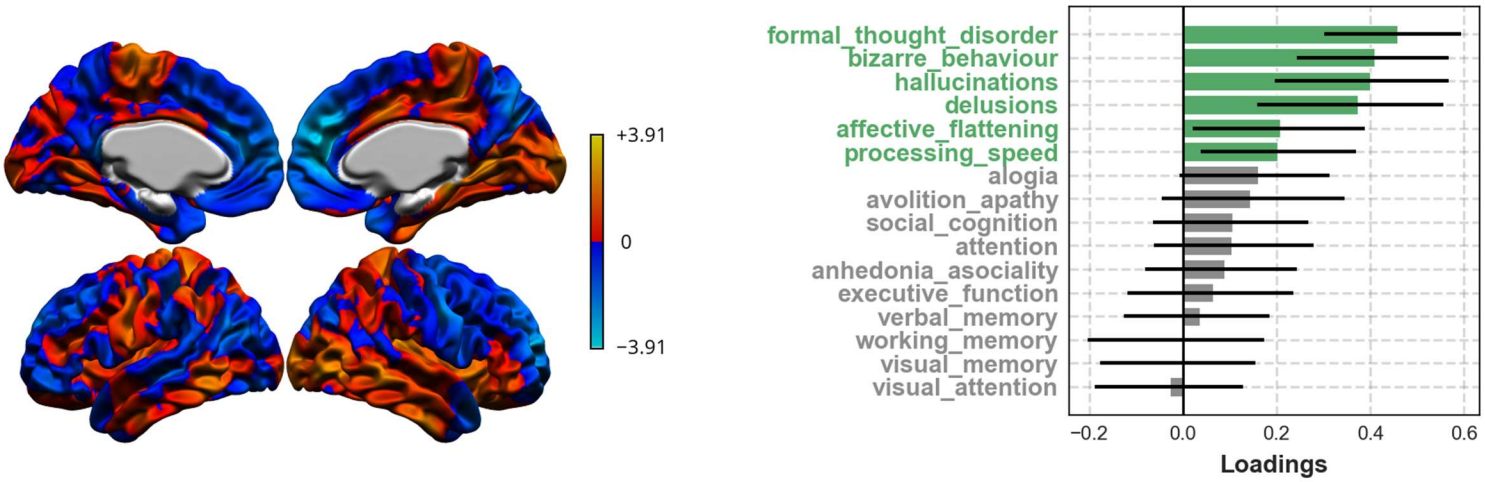


**Figure S3:** LV2 in our psychosis spectrum sample, showing bootstrap ratios for brain variables and loadings for behavioural variables, respectively. Combined, the LV mainly reflects thinner prefrontal cortex related to positive and disorganized symptoms.

**S2. Simulated data**

As we binarized each permutation test into "pass" or "fail" above, we sought to more precisely determine how the rotations impacted the null distributions in our simulated data analysis. Accordingly, for a given singular value $\sigma$ and its corresponding null distribution with mean $\bar{x}$ and standard deviation $s$, we calculated the z-value of $\sigma$ within the null distribution as:

$$z = \frac{\sigma - \bar{x}}{s}$$

Results for each simulated data analysis are shown in *Supplementary Fig. S4*.

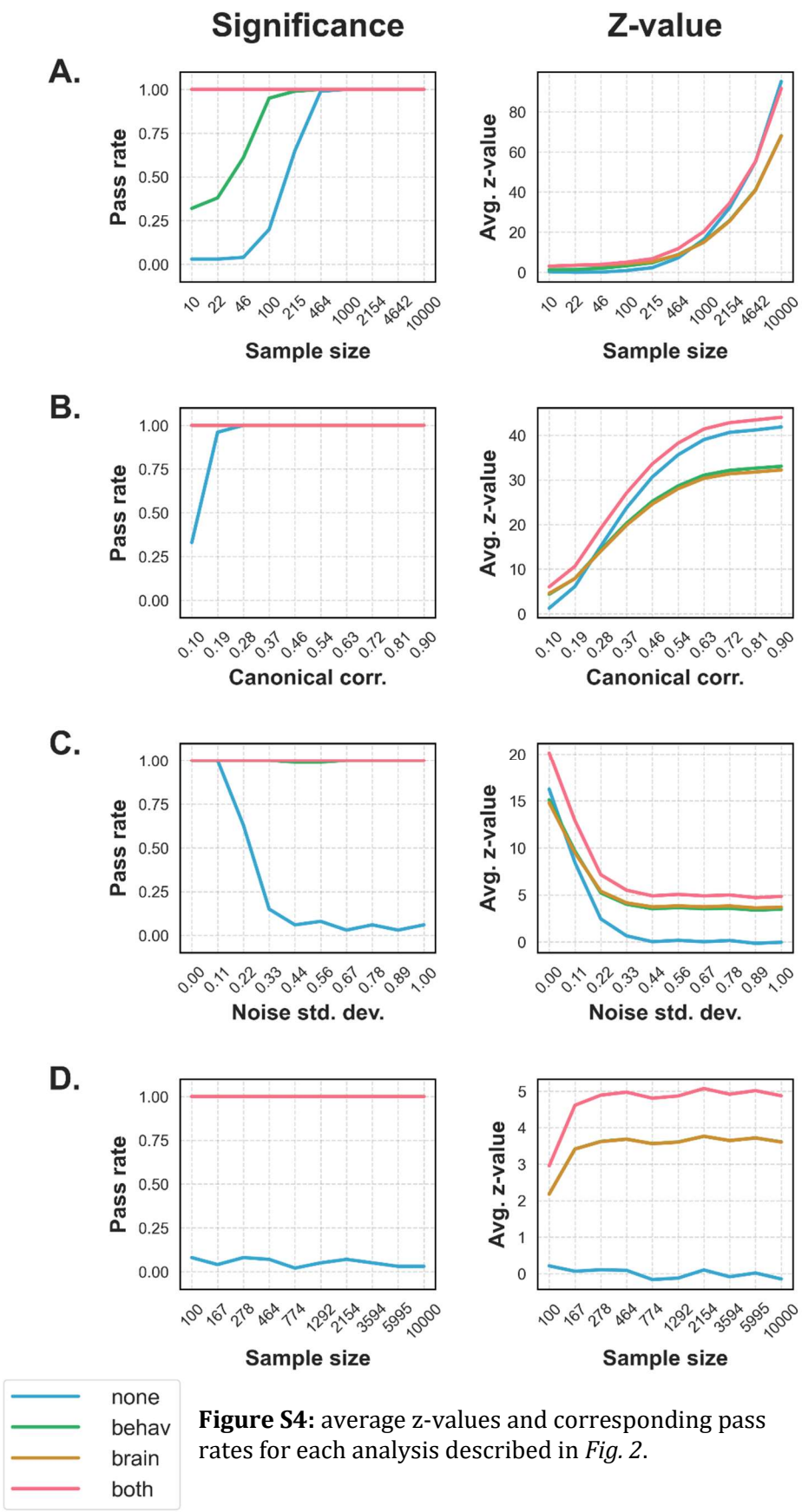


**Figure S4:** average z-values and corresponding pass rates for each analysis described in *Fig. 2*.

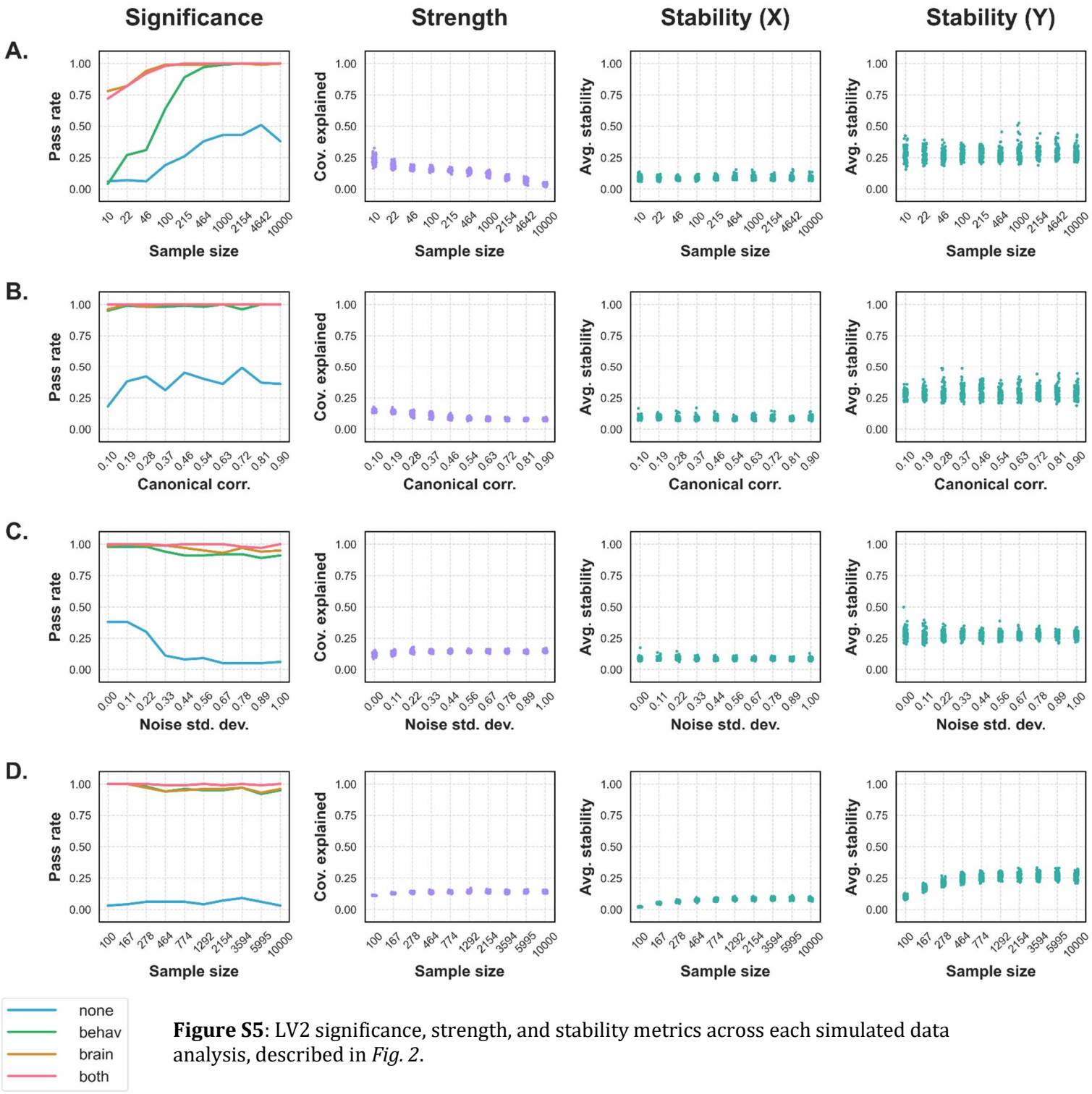


**Figure S5**: LV2 significance, strength, and stability metrics across each simulated data analysis, described in *Fig. 2*.

Importantly, LV2 was not simulated in any analysis. Accordingly, LV2 was nearly always unstable, explained a negligible amount of covariance, and consistently failed to pass unrotated permutation tests. However, rotated permutation tests consistently considered LV2 significant, suggesting that they may be sensitive to sampling variability alone.

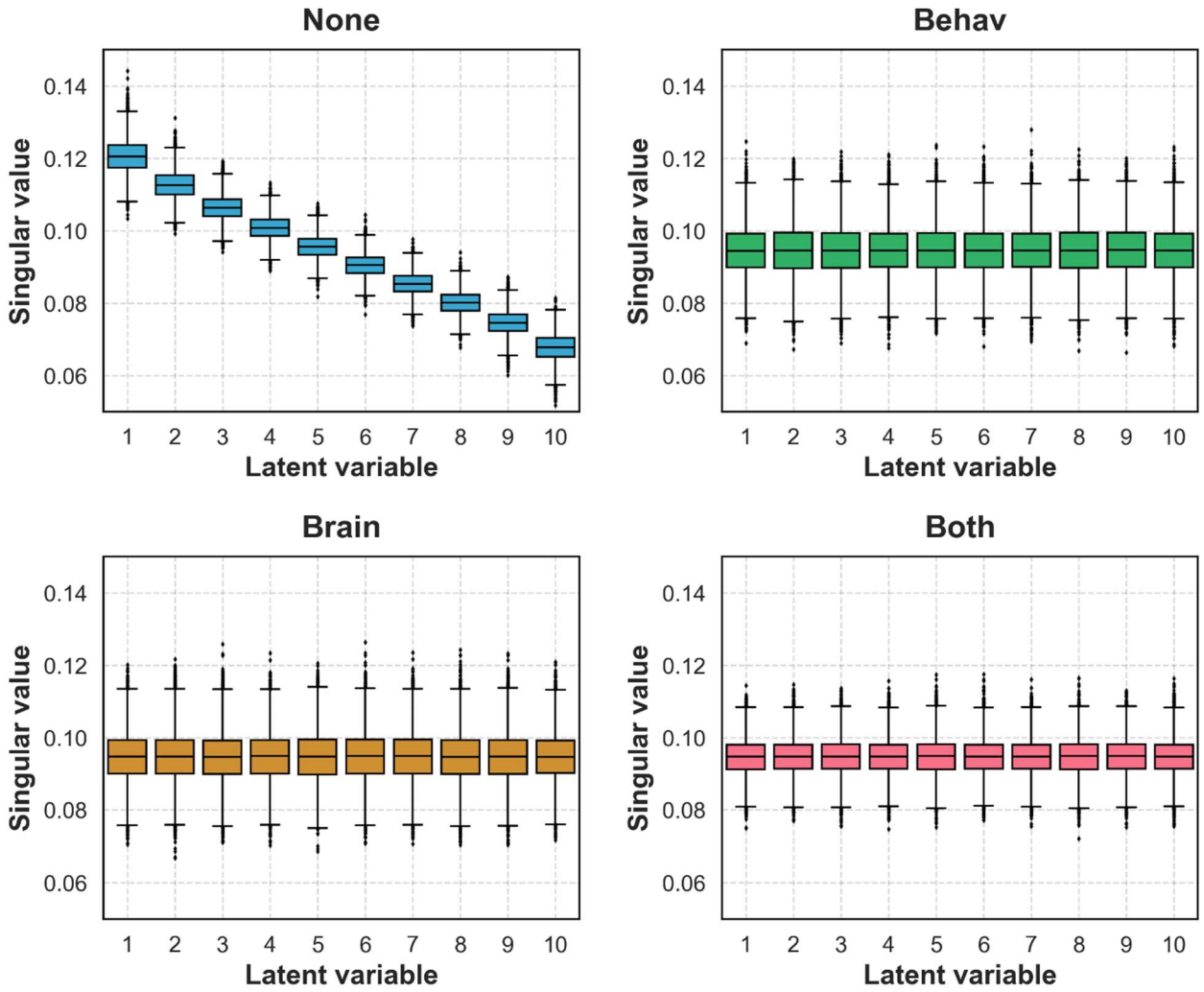


**Figure S6***:* null distributions by latent variable for a randomly generated and whitened dataset ($N$ = 10 000), plotted separately for each permutation test method. When no rotation was applied, the means of the null distributions steadily decayed, while in all rotated conditions, the distributions became visually indistinguishable. By extension, when any rotation was applied, early LVs were tested against "weaker" null distributions and late LVs were tested against "stronger" ones relative to the unrotated condition.

## S3. UK BioBank

### *S3.1. Sample*

Open access data were obtained from the UK BioBank, a population cohort of 500 000+ cognitively healthy participants aged 40 to 69 recruited across the United Kingdom between 2006 and 2010. All participants were registered with the National Health Service, lived within reasonable traveling distance of an assessment center, and provided informed written consent for the study. Full details of recruitment and other UK Biobank procedures are available online. Starting in 2014, a subset of 100 000 participants were invited to undergo multimodal magnetic resonance imaging (MRI). Of these individuals, we selected a cohort of 28 804 participants aged 45 to 82 at time of imaging who met the standards detailed below. Participants with a confounding neurological condition were excluded from the analysis.

### *S3.2. Image processing*

Data were collected with a standard Siemens Skyra 3T. Comprehensive documentation on the imaging protocol can be found online. We performed manual quality control for motion on T1-weighted (MPRAGE, 1 $mm^3$) scans prior to image processing. Scans passing quality control were input to CIVET 2.1.1 (Kim et al., 2005) to obtain regional average cortical thickness according to the Desikan-Killiany atlas.

### *S3.3. Lifestyle variables*

Behavioral, demographic, and health-related data were collected from participants at each visit to the assessment center, and here, preference was given to answers provided during the imaging visit. 38 lifestyle risk factors were selected and analyzed for a separate study, and participants were excluded if they were missing measurements for at least 10 of these variables. All further missing data was imputed using the missForest package in R. For the present study, we selected a subset of 17 of these variables with variance of at least 1 across the sample. Each is listed and described below. Questionnaire details are available online.

- *Age*: at imaging visit.
- *Alcohol intake frequency*: provided on a 5-point scale, with answer options ranging from "Never" to "Daily or almost daily".
- *Age of completing full time education*: coded as 21 for any participant who indicated having a tertiary degree (log-transformed).
- *Body-mass index*: calculated from participant height and weight measured at the assessment centre, in $kg/m^2$.
- *Days per week of walking 10+ minutes*
- *Days per week of moderate physical activity*
- *Days per week of vigorous physical activity*
- *Average total household income*: on a 5-point scale from "Less than 18 000" to "Greater than 100 000", in £ before tax.
- *Sleep duration*: average, in hours.

- *Past tobacco smoking*: provided on a 4-point scale, ranging from “I have never smoked” to “Smoked on most or all days”.
- *Pack years of smoking:* number of cigarettes smoked per day, divided by 20, and multiplied by the number of years spent smoking (log-transformed).
- *Ability to confide:* provided on a 6-point scale, ranging from “Never or almost never” to “Daily or almost daily”.
- *Frequency of family/friend visits*: provided on a 6-point scale, ranging from “Never or almost never” to “Daily or almost daily”.
- *People in household*: provided on a 4-point scale, ranging from “Only me” to “Five or more”.
- *Work/job satisfaction*: provided on a 7-point scale, with the lowest value corresponding to “I am not employed”, and subsequent values ranging from “Extremely unhappy” to “Extremely happy”.
- *Systolic blood pressure*: measured manually or automatically by registered nurses using an Omron 705 IT electronic blood pressure monitor.
- *Diastolic blood pressure*: as above.

### *S3.4. Figures*

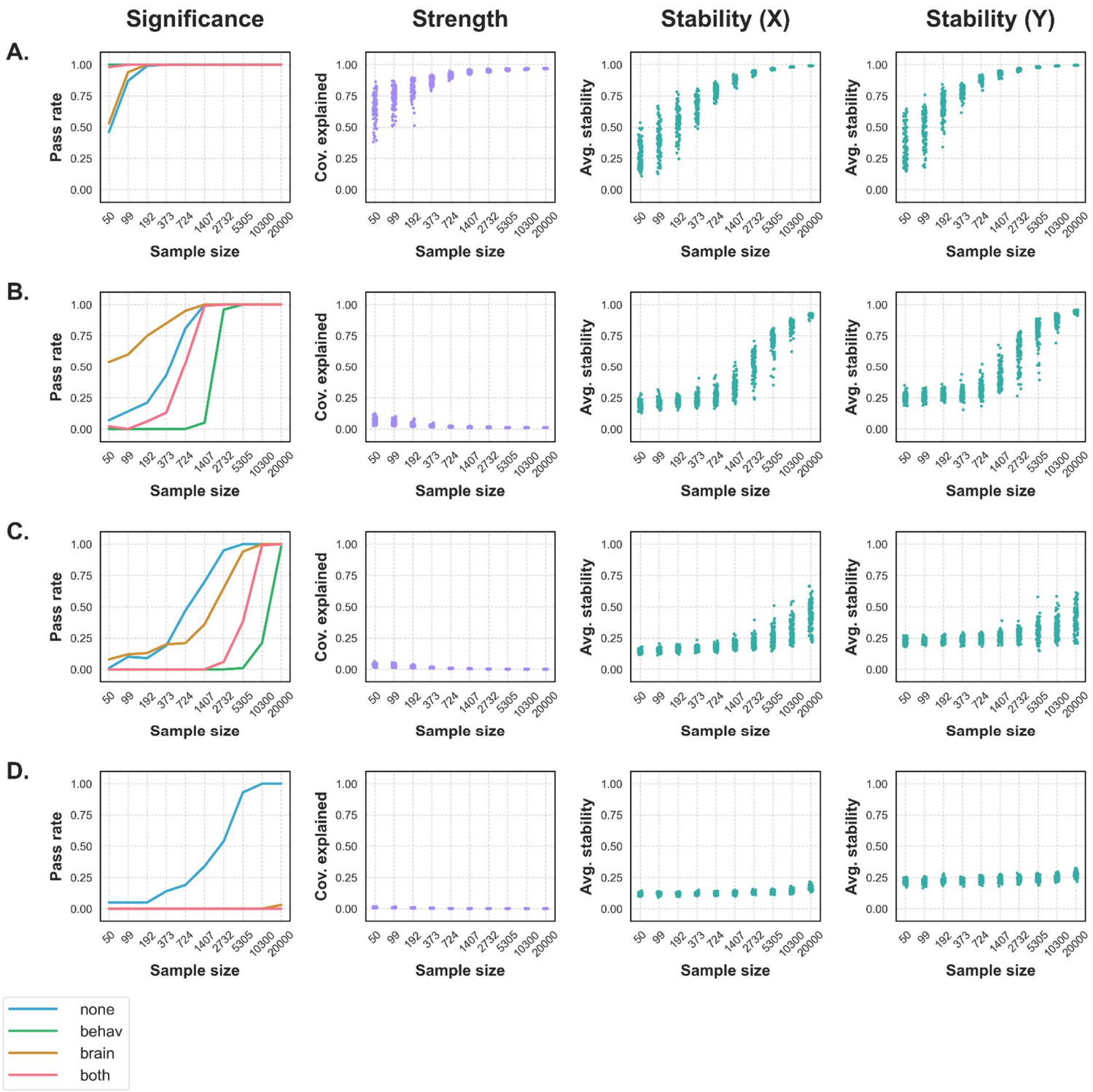
Significance
Strength
Stability (X)
Stability (Y)
A.
B.
C.
D.
Pass rate
Cov. explained
Avg. stability
Sample size
0.00
0.25
0.50
0.75
1.00
50
99
192
373
724
1407
2732
5305
10300
20000
none
behav
brain
both

**Figure S7**: LV1, 3 5, and 10 significance, strength, and stability metrics across subsamples of our UK BioBank sample. Columns are described in *Fig. 2.*

**A. (LV1)** LV1 was consistently considered significant, strong, and stable at most sample sizes.

**B. (LV3)** The permutation tests disagreed regarding LV3 significance, except at the largest sample sizes. Notably, the unrotated permutation test is not the most conservative method. While LV3 did not explain much covariance, it was considered stable in large samples.

**C. (LV5)** The permutation tests again disagreed on LV significance, with the unrotated test now the most liberal method. Meanwhile, LV5 was not consistently considered strong or stable.

**D. (LV10)** Only the unrotated test considered LV10 significant, and only in the largest samples. LV10 was never considered strong or stable.

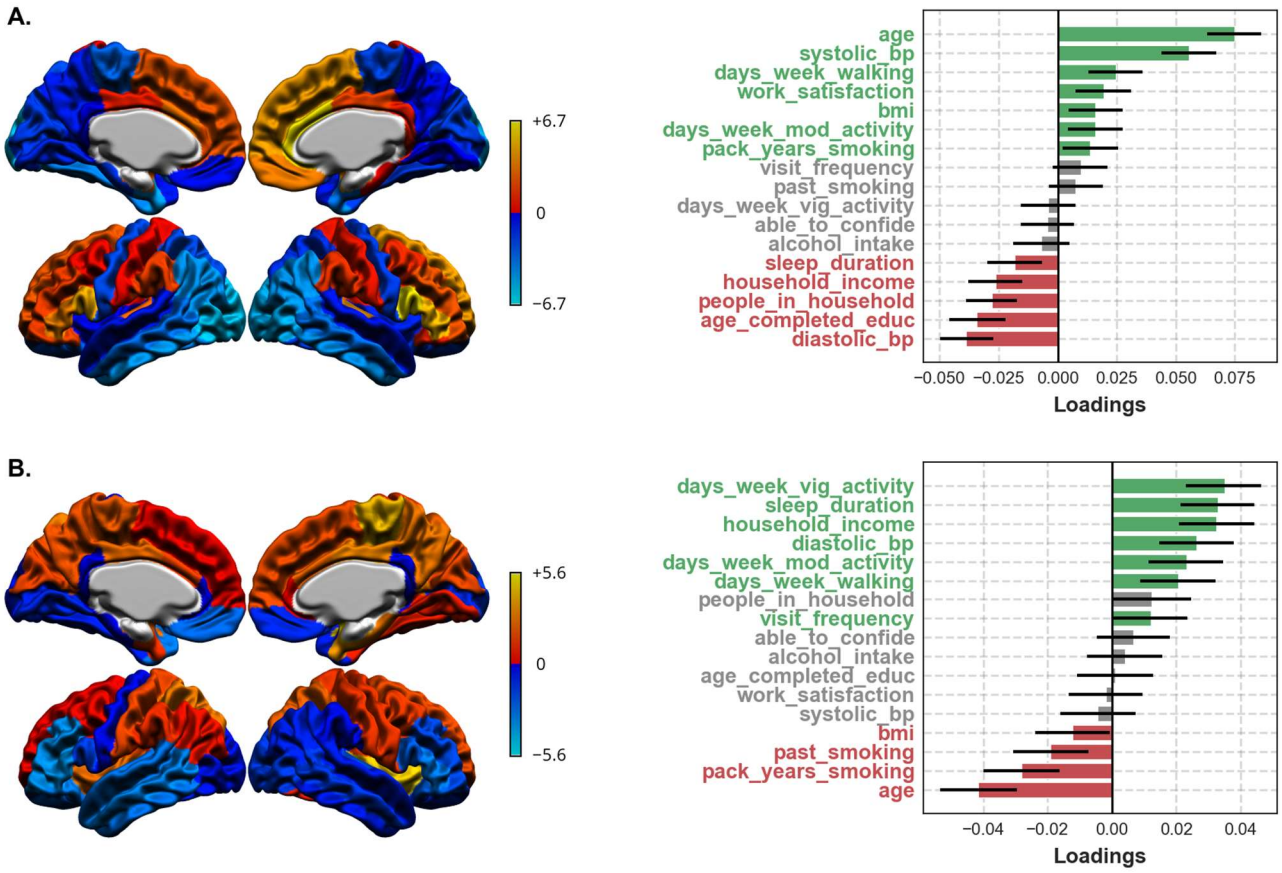


**Figure S8**: LVs 5-6 in the full UKBB sample.

**A. (LV5)** Combined, the LV reflects a cortical thickness pattern related most strongly to age, which was already a dominant contributor to LV1.

**B. (LV6)** Together, no dominant sets of brain or behavioural variables were observed.

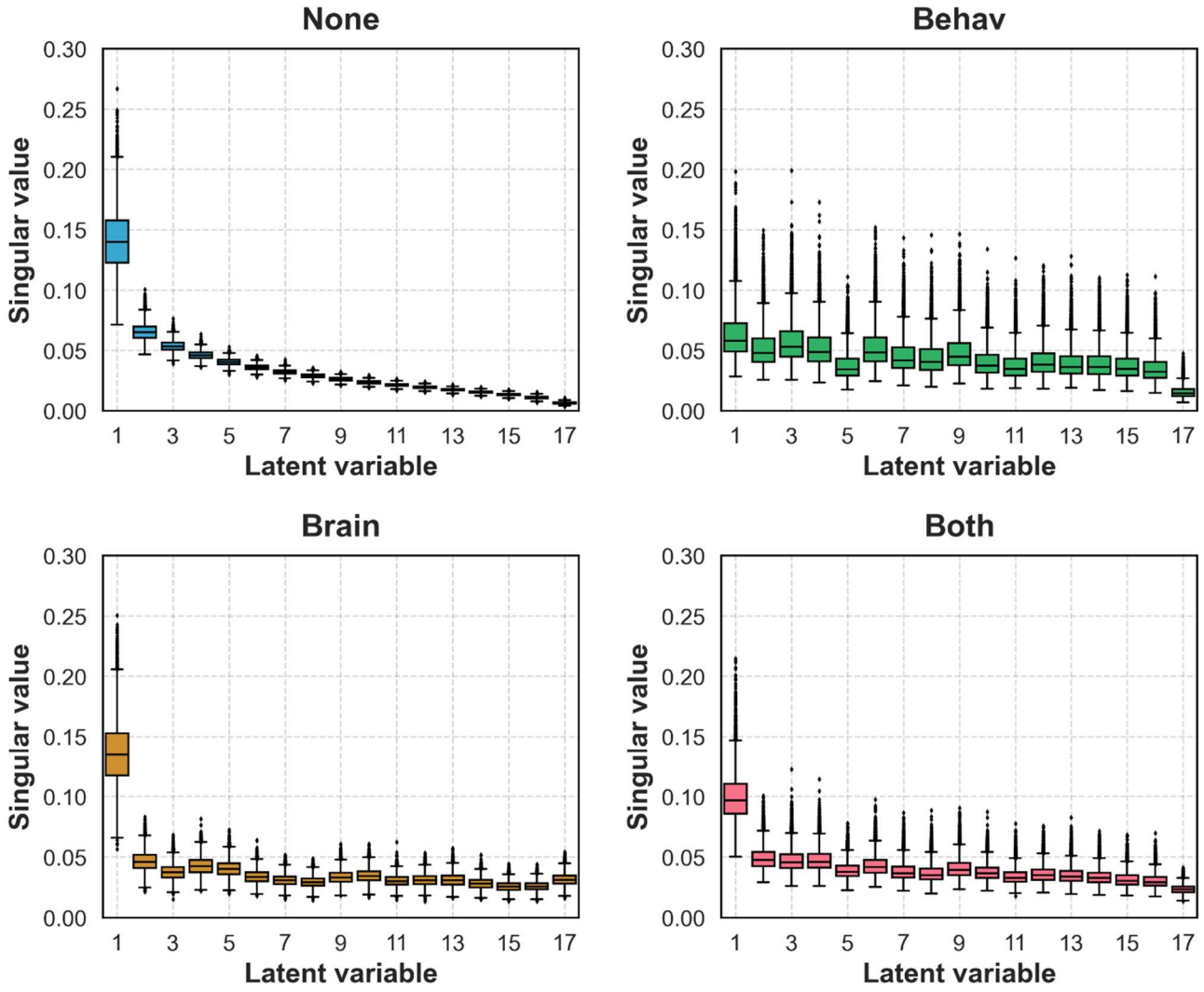


**Figure S9:** null distributions by latent variable for the full UK BioBank sample, plotted separately for each permutation test method. Qualitatively, later null distributions tended to have a higher mean and a greater spread of values after any rotation was applied. In other words, late null distributions appeared to "gain" covariance at the expense of early ones following a rotation. This effect was especially pronounced when the behavioural component was rotated. By extension, when any rotation was applied, early LVs were tested against "weaker" null distributions, and late LVs were tested against "stronger" ones – demonstrating why only the unrotated tests deemed late LVs significant.